

The Tulsa Processor: A Dual Core Large Shared-Cache Intel® Xeon™ Processor 7000 Sequence for the MP Server Market Segment

Jeffrey D. Gilbert
Stephen H. Hunt
Daniel Gunadi
Ganapati Srinivas

Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

Intel may make changes to specifications and product descriptions at any time, without notice.

All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit <http://www.intel.com/performance/resources/limits.htm> or call (U.S.) 1-800-628-8686 or 1-916-356-3104.

Relative performance for each benchmark is calculated by taking the actual benchmark result for the first platform tested and assigning it a value of 1.0 as a baseline. Relative performance for the remaining platforms tested was calculated by dividing the actual benchmark result for the baseline platform into each of the specific benchmark results of each of the other platforms and assigning them a relative performance number that correlates with the performance improvements reported.

64-bit Intel® Xeon™ processors with Intel® EM64T requires a computer system with a processor, chipset, BIOS, OS, device drivers and applications enabled for Intel EM64T. Processor will not operate (including 32-bit operation) without an Intel EM64T-enabled BIOS. Performance will vary depending on your hardware and software configurations. Intel EM64T-enabled OS, BIOS, device drivers and applications may not be available. Check with your vendor for more information.

SPECint2000 and SPECfp2000 benchmark tests reflect the performance of the microprocessor, memory architecture and compiler of a computer system on compute-intensive, 32-bit applications. SPEC benchmark tests results for Intel microprocessors are determined using particular, well-configured systems. These results may or may not reflect the relative performance of Intel microprocessor in systems with different hardware or software designs or configurations (including compilers). Buyers should consult other sources of information, including system benchmarks; to evaluate the performance of systems they are considering purchasing.

Intel, Xeon, Itanium, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2006 Intel Corporation.



Agenda and Anti-Agenda

Agenda

- The market conditions surrounding and informing Tulsa's definition
- Guidelines for making high performance OLTP server processors
- Selecting amongst the options for FSB multi-core processors
- Tulsa's implementation experience
- Tulsa's performance results
- Concluding remarks

What this talk is not about (see references slides)

- The Intel® 64 ISA
- The Netburst® Microarchitecture
- Intel's 65 nm process technology

Tulsa Feature Overview

Large shared 16M L3 cache

- Provides 70% (and more) performance boost to applications in *existing* platforms

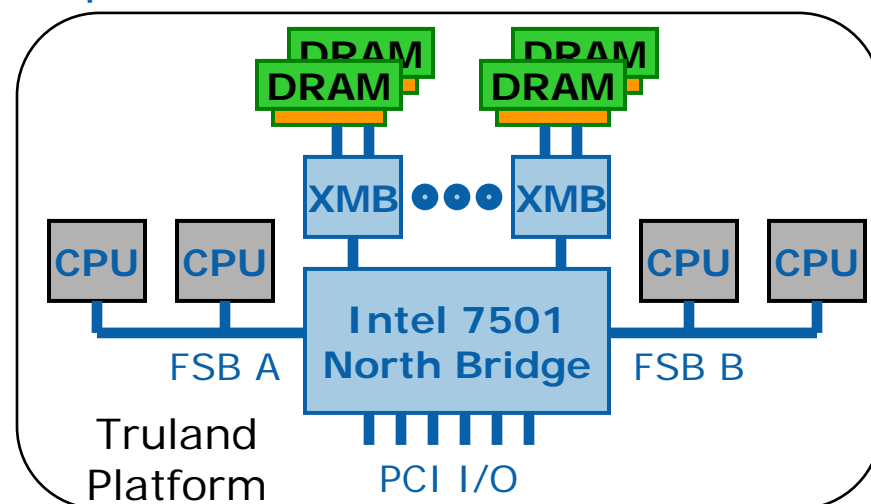
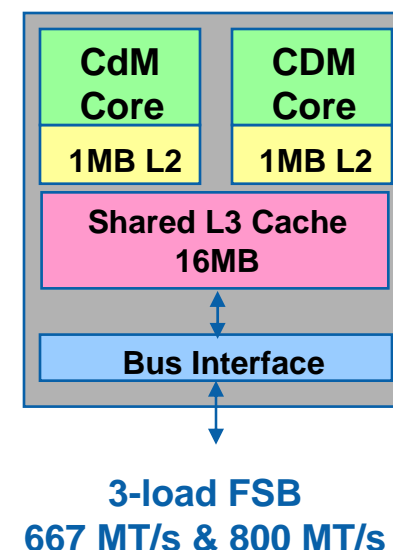
Two Netburst® (a.k.a. Pentium® 4) cores on a single die targeting 3.4 GHz core frequency

- Four threads per processor with HT enabled on each core

Designed for existing 667/800 MT/s FSB platforms

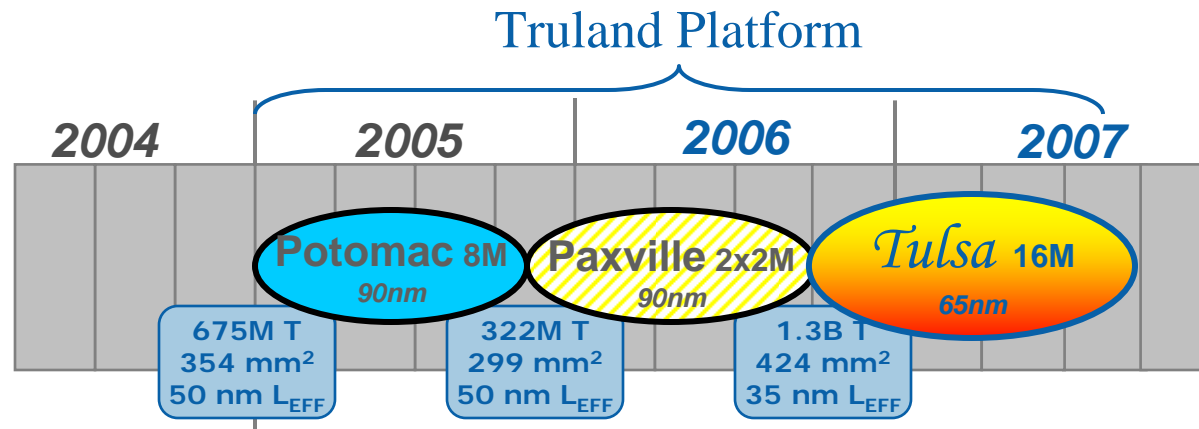
Based on 65nm process technology

- 150 and 95 Watt SKUs
- Intel® Cache Safe Technologies for improved RAS
- Virtualization technology for improved robustness
- SMBus system management interface for better manageability



Compelling features enabling a performance boost, improved RAS, and manageability

Where does Tulsa fit?



- General MP market segment expectation for performance growth: 40% to 65% “CAGR” (compound annual growth rate)
- An MP platform has to last for 30 to 36 months
 - OEM validation and marketing costs are amortized over that lifetime
 - Socket compatible processors have to boost performance 2x to 3x
- Truland (with its Twin Castle central agent) spanned the single to dual core Xeon MP processor transition – a huge performance range

How do you make a fast MP Server CPU?

- Pluses and minuses of designing a processor for an existing platform
 - The platform is stable and reliable
 - The system interface is fixed
 - The power envelope is fixed
 - The memory subsystem and I/O subsystem are defined
- Optimize for the target applications
 - Examples: Transaction Processing and Enterprise Resource Planning
- The components and tools at hand
 - 65 nm NetBurst core (internally named “Cedar Mill”)
 - New core or radical changes were not schedule or resource feasible
 - Cedar Mill brought power efficiency and reliability benefits
 - Silicon technology and capacity for a large cache in addition to two cores
 - Experienced server CPU design team



Optimizing for OLTP - I

- From the processor perspective, a single transaction can be described as:

$$t_{\text{TRAN}} = \text{PL} \times \text{TPI} + \text{Mem} \times \text{Mem_Lat}$$

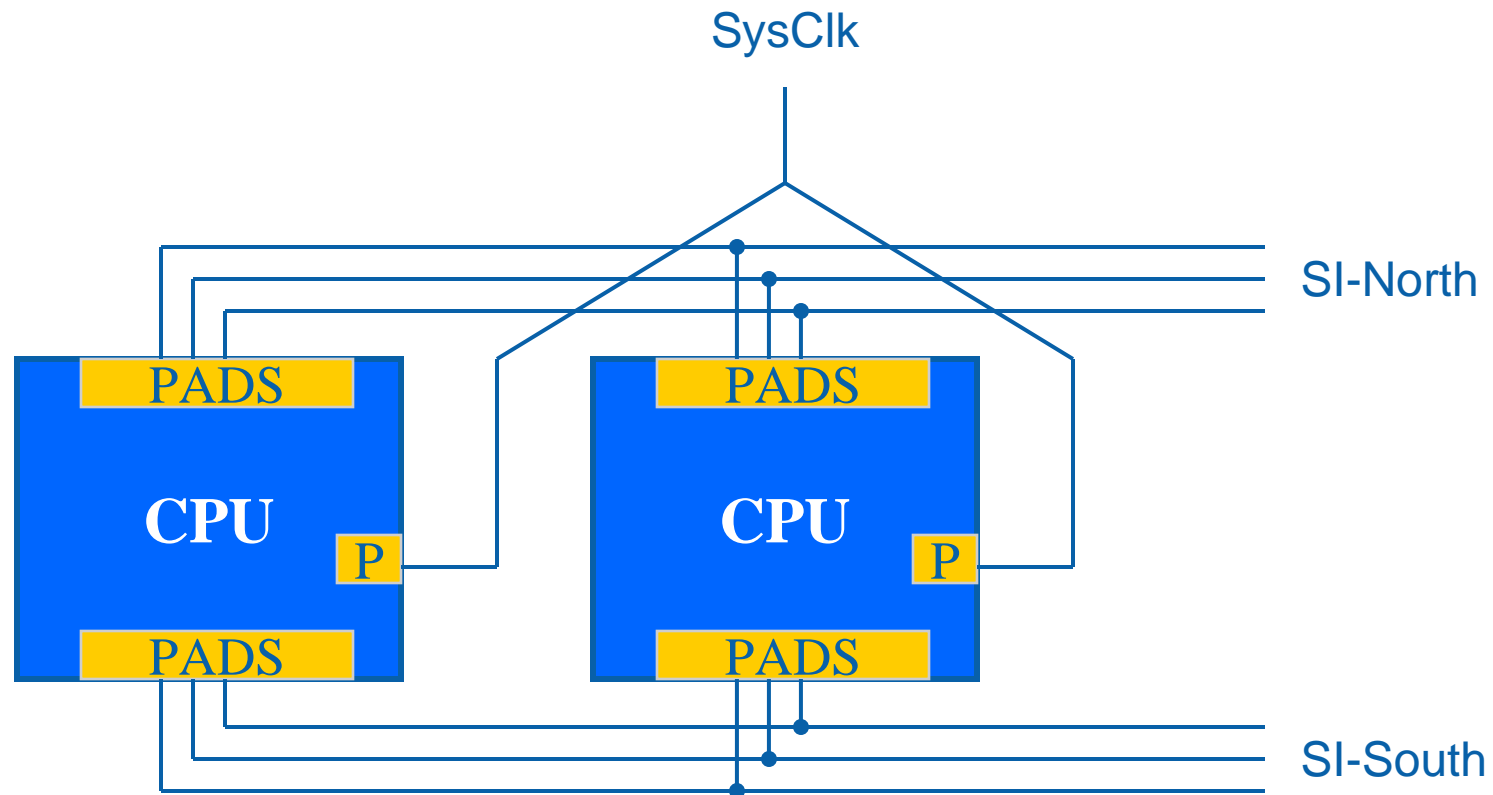
- t_{TRAN} is the time for a thread to complete a transaction
- PL is the “Path Length” – or number of instructions per transaction
 - This number is architecture and micro-architecture dependent
 - It also varies by performance level
 - Linear approximation derived from platform experiments
- TPI is the average “Time per Instruction”
 - Derived by measuring traces of the application workload
- Mem is the number of serializing memory fetches per transaction
 - Derived from system measurement
- Mem_Lat is the effective memory latency
 - “effective” is the key here
 - System activity affects memory latency, too
 - Overlapped execution helps!

Optimizing for OLTP - II

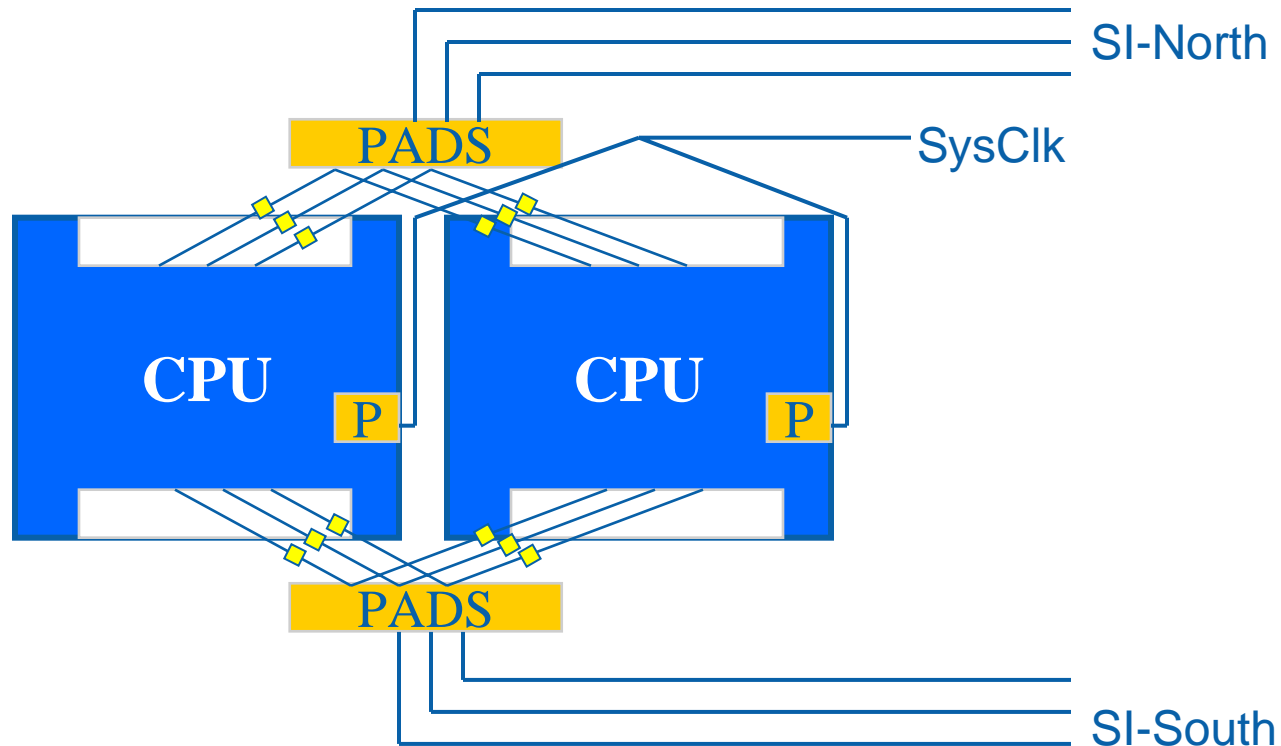
- PL is pretty much a given (but you can work with the compiler folks)
- TPI can be influenced ...
 - ... by some microarchitectural features such as buffer and cache size
 - ... and by the core's operating frequency
- Mem can also be influenced, too ...
 - ... by the size of the core cache (fewer core cache misses)
- The real leverage point is "Mem_Lat"
 - With an added level of cache hierarchy there's a opportunity here
 - Important parameters contributing to overall memory latency
 - Cache hierarchy hit time – both to hit/miss and data return
 - Cache hierarchy miss time to system fetch
 - Cache hierarchy hit rate
 - In a multi-core design: cross-core snoop time
- More threads, of course can directly scale the performance – a first order effect.

Three Paths to Multi-Core Designs

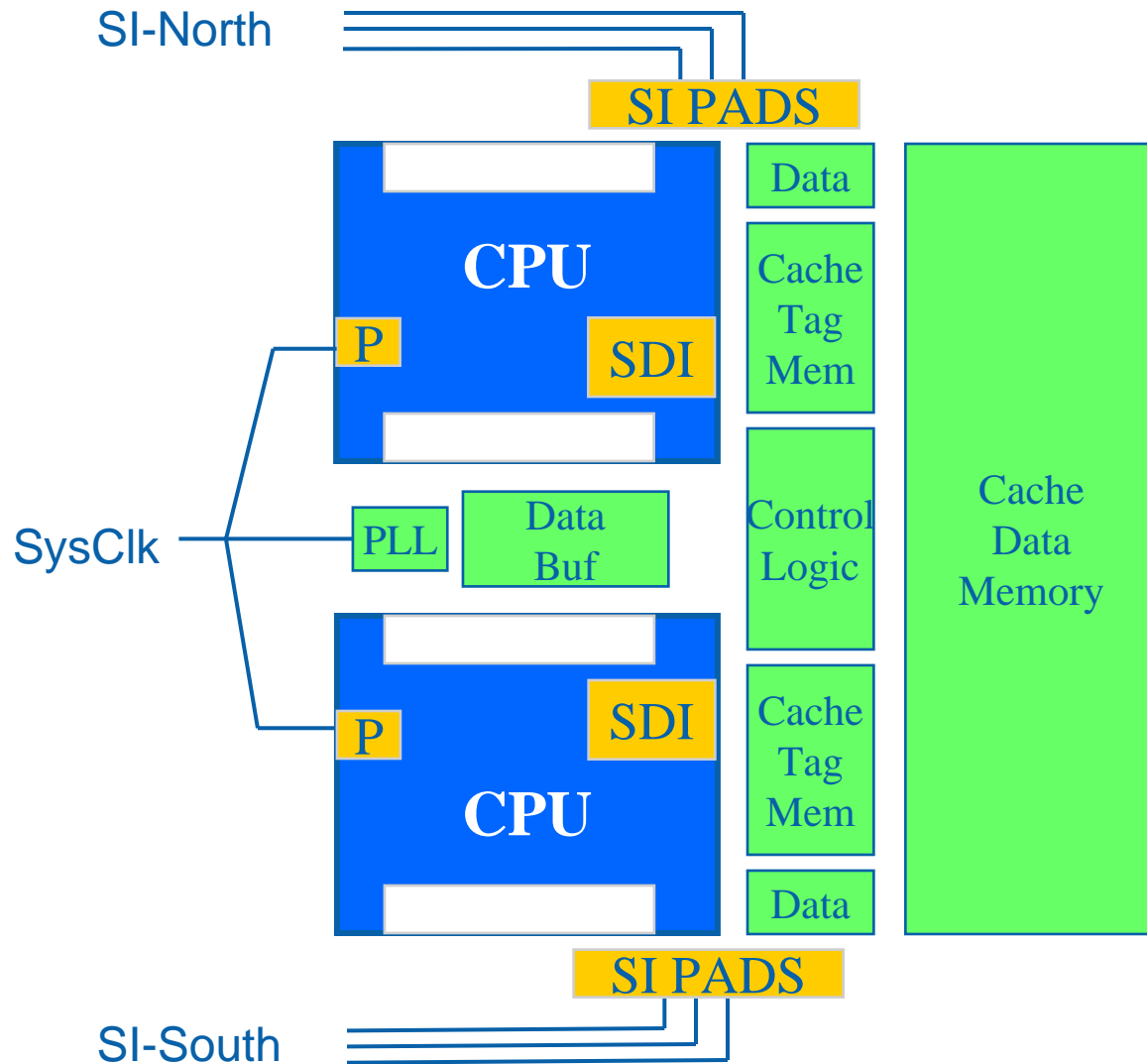
Multi-Core Option #1: Join at the System Interface Pads



Multi-Core Option #2: Join at the Pad Digital Interface



Multi-Core Options #3: Integrated UnCore



Other Aspects of Options for Multiple Cores

- The three approaches all increase the number of threads
- Options 1 and 2 are almost the same but the choice between them may be dictated by system topologies
 - These options rapidly provide the benefits of multi-core processors
 - Cache sharing only through the FSB
 - Seen in Xeon, Itanium, and even non-Intel multi-core CPUs
- Options 3 – efficient sharing of the outer level cache (“Last Level Cache” or LLC) offers performance scaling beyond thread counts
 - The efficiency of core-to-unCore communication is critical
 - Core fetches that miss the LLC have added latency to the system interface
 - A performance increase occurs when the saving from servicing some core fetches out of the LLC outweigh the added latency for LLC misses
 - There may also be system interface queue latency benefit from the LLC

The Tulsa Engineering Experience - I

- The Potomac project used its core's FSB as the on-die interconnect
 - At the time, the design simplification was viewed as an acceptable tradeoff against the latency/performance consequences
 - The latency benefit of Potomac's on-die LLC was realized but muted somewhat by the FSB protocol's inherent latency
- There was some trepidation about replacing the FSB logic with a new, on-die interface
 - The FSB logic is well understood – having perhaps a dozen incarnations
 - The “Simple Direct Interface” (SDI) replacement logic promised better performance but was still in its design phase at decision time
 - A shared cache deviated from Xeon's traditional approach of adding more cache or another level of cache to the core's cache management logic
 - The latency and cache efficiency benefit was so compelling that the risk was judged appropriate

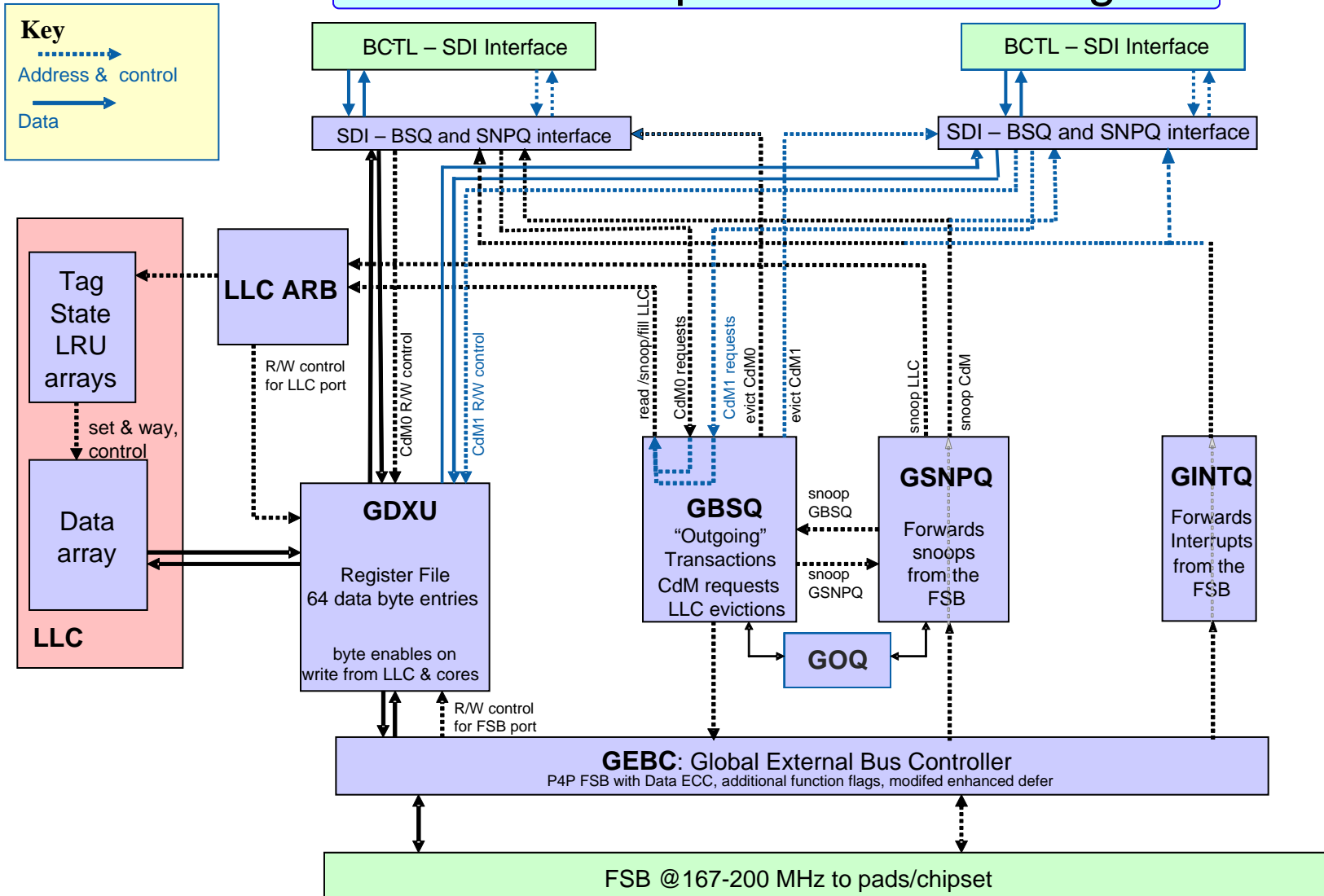
The Tulsa Engineering Experience - II

- Previous work suggested memory ordering and transaction conflicts would be problem areas
 - The unCore must act like a bridge between the system interconnect and the cores – correctly conveying global observation
 - Multiple transaction in process can reference the same cache line requiring conflict detection and handling
 - Optimizing performance – by overlapping and re-ordering operations – generally works against ordering and puts pressure on conflict handling

The Tulsa Engineering Experience - III

- As it turned out ...
 - The efforts to implement the interface was considerably smaller than first estimated and was carried out relatively smoothly
 - Clocking – in general – and Intel's established technique for power management by frequency/voltage scaling were more difficult to implement than anticipated
 - The flexibility by transcending some FSB protocol limits accelerated performance and simplified conflict resolution
 - Greater parallelism of SDI by removing some sequencing requirements
 - Removed completion restriction on capacity eviction operations
 - Memory ordering as reflected in the cores required careful attention, but earlier work proved effective with Tulsa
 - Cache replacement policies can play a dramatic part in LLC efficiency
 - Changing Tulsa's LRU update policy on core cache capacity evictions yielded a double-digit performance benefit for OLTP applications

Tulsa unCore μ Architectural Diagram



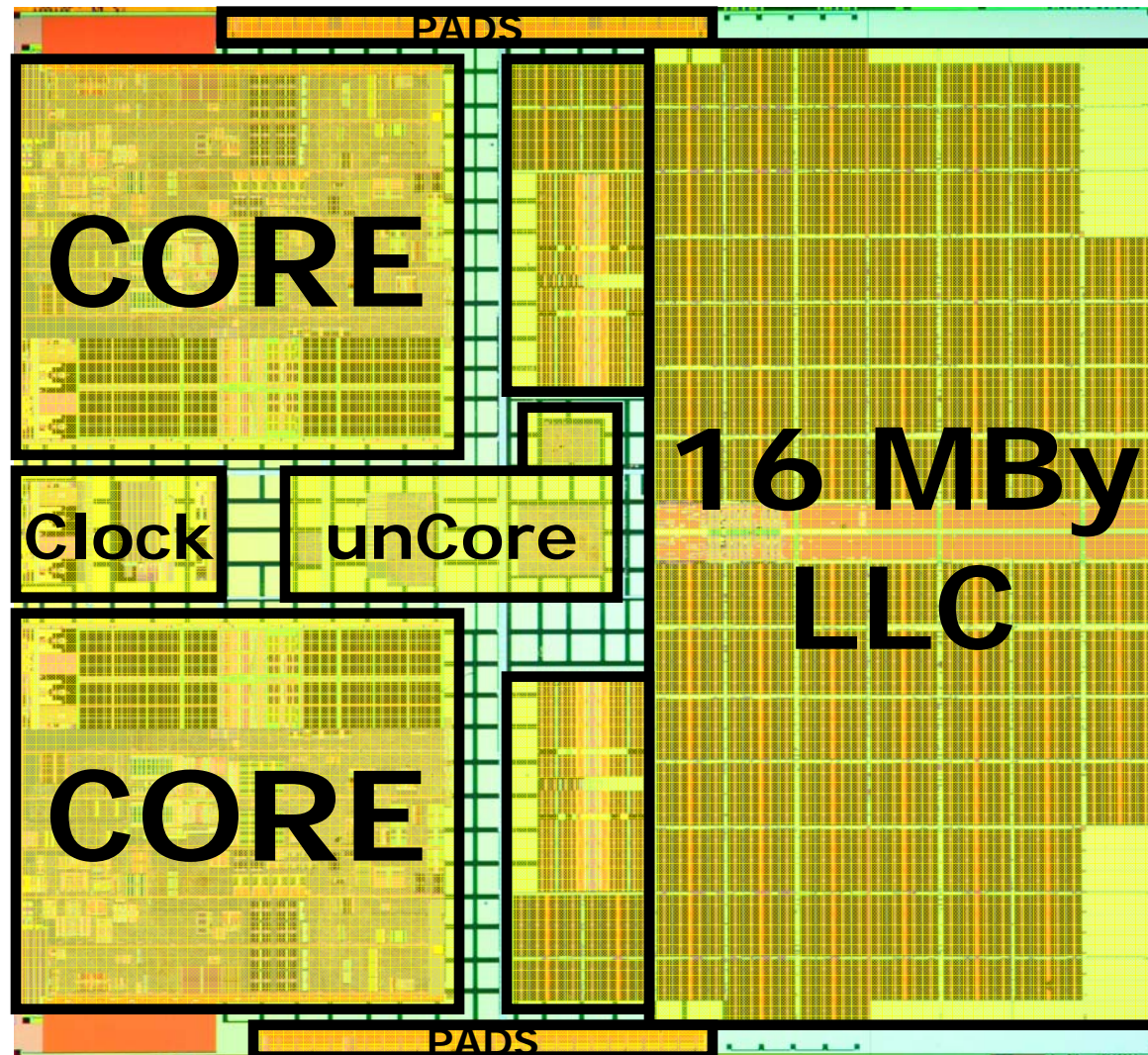
Tulsa Microarchitectural Diagram Notes:

Unit Summaries with Rates and Bandwidths

- GBSQ holds transactions issued by the cores and LLC capacity evictions
 - GSNPQ sequences FSB snoops to the LLC and cores
 - GINTQ conveys FSB interrupt transactions to the cores
 - GDXU is a cache line register file (64 bytes each)
 - GOQ preserves FSB cache line ownership transfer ordering
 - LLC ARB accepts LLC requests
 - GEBC convey FSB requests into and out of Tulsa's unCore
-

- 200 MHz / 800 MT/s FSB: 100 M requests/sec, 6.4 GBy/sec
- LLC: 280 M requests/sec, 18.1 GBy/sec rd, 18.1 GBy/sec wr
- GDXU: 1.7+ G reads/sec, 54+ GBy/sec; 1.7+ G writes/sec, 54+ GBy/sec
- SDI: 425 M request/sec, 13.6 GBy/sec rd, 13.6 GBy/sec wr

Tulsa Die Shot



INTEL Corporation

19 21 Aug 06 The Tulsa Processor
JDG



The Tulsa Performance Notes - I

- Clock generation and clock domain crossing turned out to be more expensive than originally expected
 - Nearly 20% of the unCore's LLC hit latency is a result of clock crossing
 - At least one snoop stall is traceable to bus-to-cache clock crossing
 - No obvious engineering alternative with multiple clock domains
- The NetBurst replay loop time was a challenge to plan for
 - The core cache access time is synchronized with the replay loop time
 - The quantized thresholds make it difficult for a less tightly coupled LLC to optimize its data access time
 - If data delivery misses a replayed micro-Op by even one core clock, data latency stretches to the next replay point
 - Over the course of the design, the margin initially provided was exceeded making Tulsa suffer an extra replay interval for LLC hits

The Tulsa Performance Notes - II

- Fortunately, a large shared cache has many salutary effects
 - 2005 versions of OLTP applications (in EM64T) have considerably larger code footprints – playing to the strength of a big LLC
 - The FSB traffic foregone by LLC completion of core fetches is critical to achieving higher performance levels by providing “room” for I/O traffic
 - The performance increase achieved by an LLC provides a significant performance-per-unit-power (power efficiency) boost
 - Cross-core data sharing is a relatively small component of overall traffic, but every bit helps

Tulsa Performance Results - I

- Tulsa's optimization for OLTP – as shown by the marquee OLTP performance – exhibit a nearly 70% improvement over the previous generation using the same platform
 - Literally a drop-in replacement to achieve this performance gain
- This is a remarkable result given ...
 - Tulsa's cores have 1 MBy mid-level cache versus the previous generation's 2 MBy mid-level cache
 - The platform has rather lengthy idle memory latency, measured on the FSB at about 150 ns (request to data delivery)
 - Tulsa itself experiences longer memory latency (see next slide)
 - The cores on Tulsa operate 13% faster than the previous generation, but OLTP performance is generally weakly correlated to core frequency)
- Tulsa is also a significantly lower power part than Paxville-MP
 - Cores operate below 65 W each – a 20% to 40% improvement over the previous process generation

Tulsa Performance Results - II

- Measured Tulsa characteristics
 - Core to mid-level cache load-to-use time: ~ 7.5 ns
 - Core to LLC cache load-to-use time: ~ 35 ns
 - The technology allows a 16 MBy cache to have < 9 ns access time
 - All of the load-to-use stages and clock crossings cost a lot
 - Core to memory (idle) load-to-use time: ~ 195 ns
 - About 15 ns longer than the previous generation
 - OLTP applications typically experience a 50% - 60% LLC hit rate

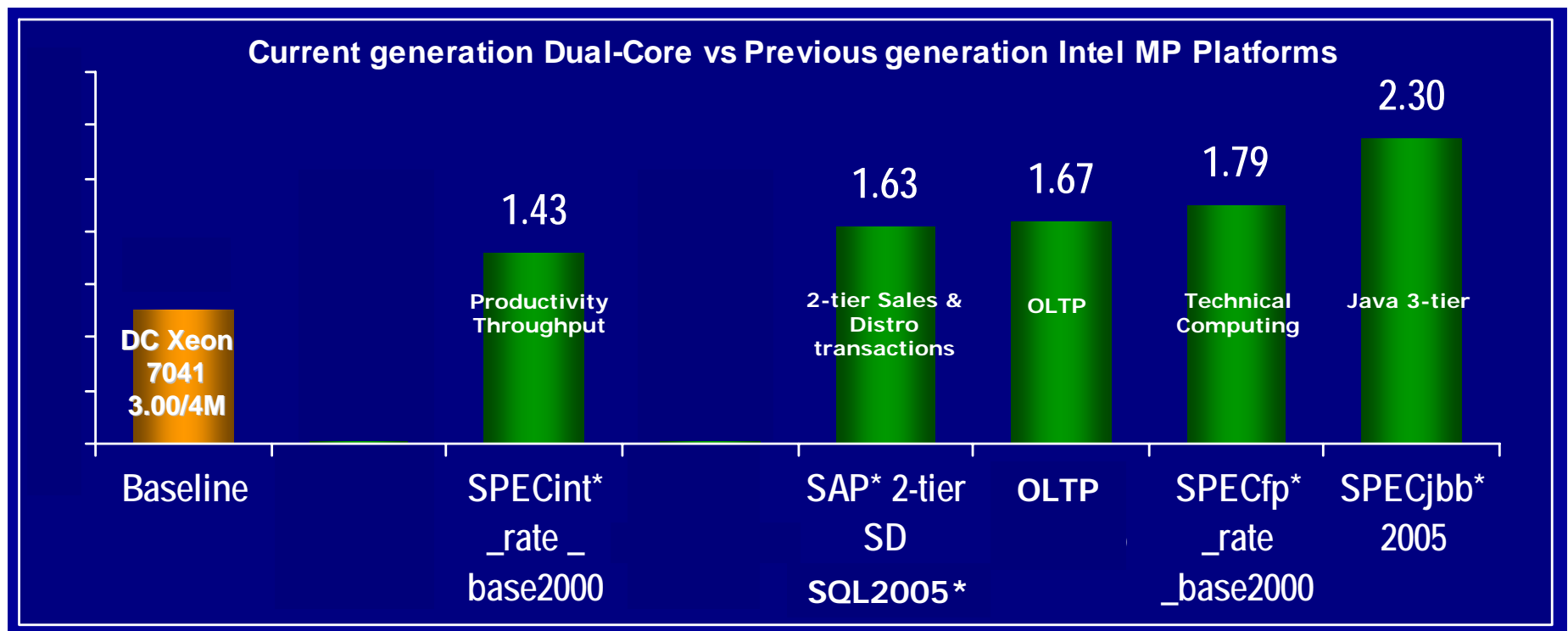
Tulsa Performance Results - III

- Tulsa's goal of reducing memory latency for 4P OLTP
 - 60% of core requests completed on die in about 35 ns
 - 40% of core requests have a 15 ns unCore propagation time added
 - Paxville experiences lengthy FSB latencies from high utilization (~90%)
 - Tulsa reduces utilization of the FSB
 - LLC's on-die completion don't reach the FSB
 - The Defer Phase sub-bus move completing off the request sub-bus
 - Approx effective latency for core transactions (4P)
 - $60\% * 35 \text{ ns} + 40\% * 240 \text{ ns} \rightarrow 117 \text{ ns}$
 - This is about 1/3 of the 4P Paxville-MP's effective memory latency
 - This comparison accounts for the latency effects of Paxville-MP's 2 MBy core caches (versus Tulsa's 1 MBy core caches)

Tulsa MP Server Application Performance

Compares Tulsa platform to prior generation Intel MP platform

- Tulsa platform: Dual-Core Intel® Xeon® Processor MP 7140M
- Baseline platform: Dual-Core Intel® Xeon® Processor MP 7041



Data Source: Publicly posted results and Intel internal measurement (July 2006). See backup for links and details.

INTEL Corporation

25 21 Aug 06 The Tulsa Processor
JDG

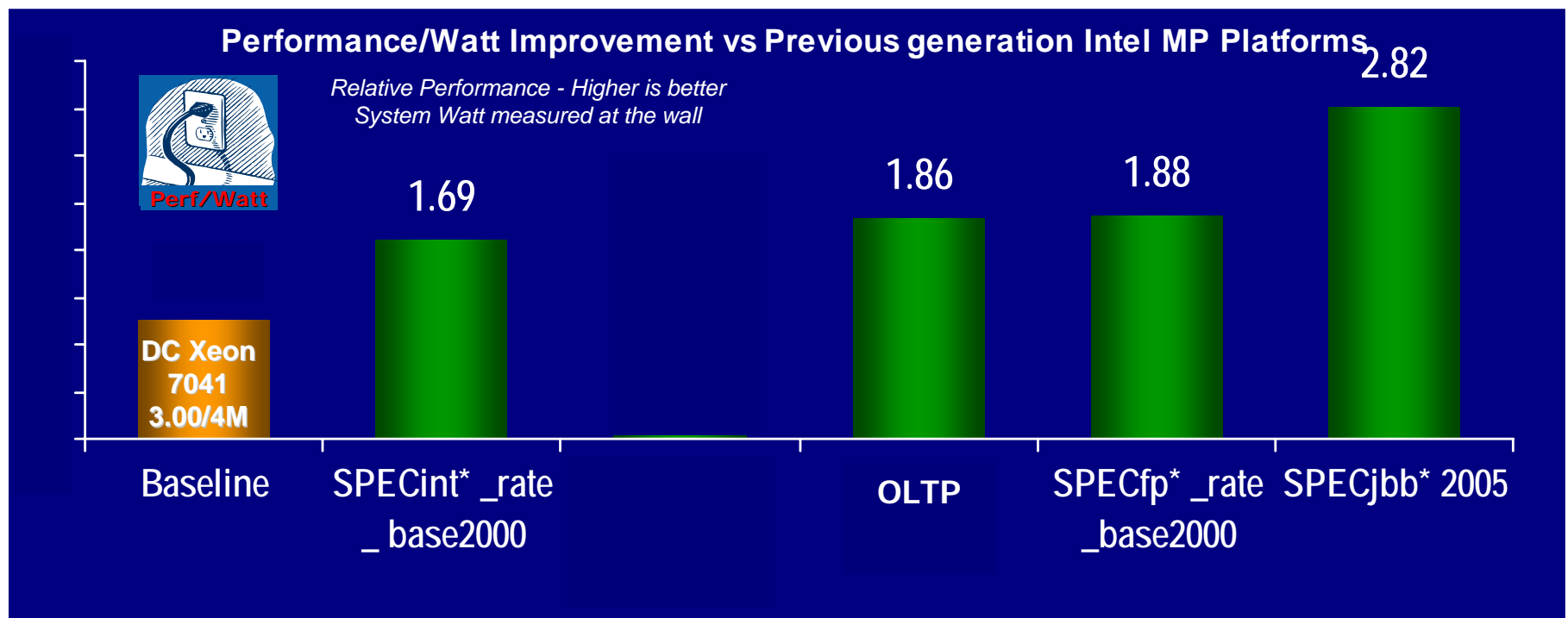
*Other names and brands may be claimed as property of others. System Configuration details and/or links to posted results in backup. Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit <http://www.intel.com/performance/resources/limits.htm> or call (U.S.) 1-800-628-8686 or 1-916-356-3104.



Tulsa Performance/Watt (PPW) Comparison

Compares Tulsa platform to prior generation Intel MP platform

- Tulsa platform: Dual-Core Intel® Xeon® Processor MP 7140M
- Baseline platform: Dual-Core Intel® Xeon® Processor MP 7041



Data Source: Publicly posted results and Intel internal measurement (July 2006). See backup for links and details.

INTEL Corporation

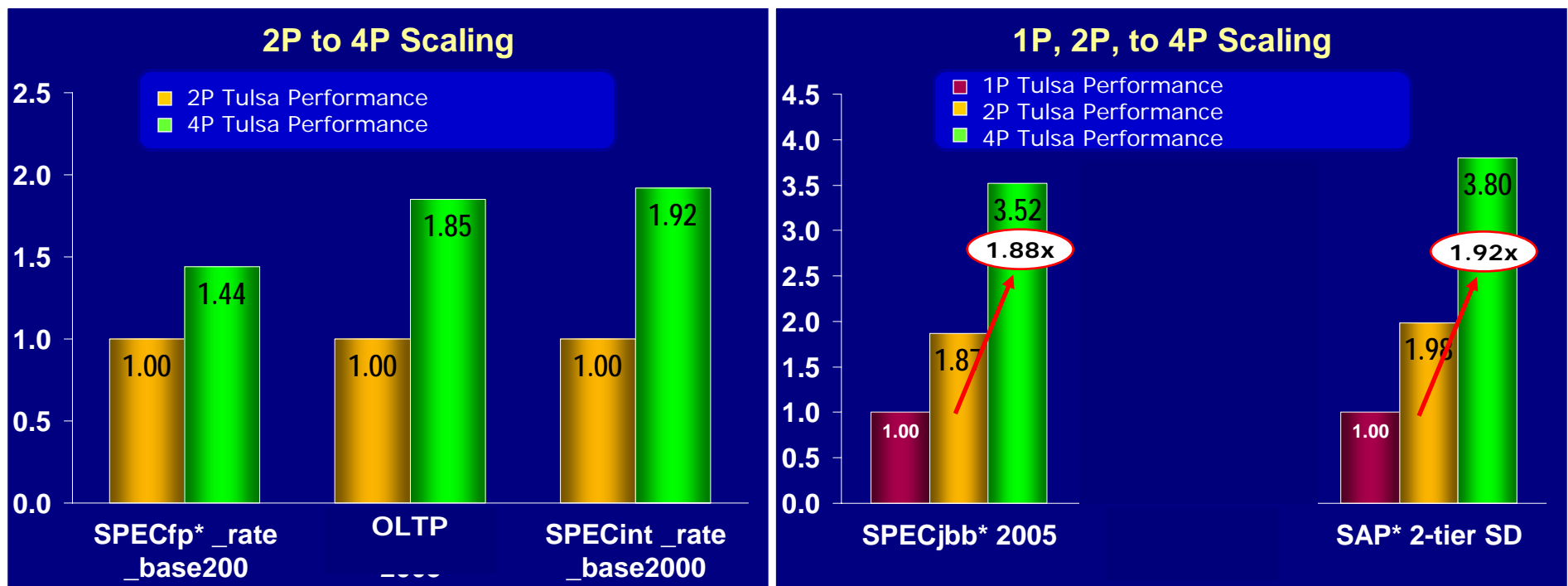
26 21 Aug 06 The Tulsa Processor
JDG

*Other names and brands may be claimed as property of others. System Configuration details and/or links to posted results in backup. Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit <http://www.intel.com/performance/resources/limits.htm> or call (U.S.) 1-800-628-8686 or 1-916-356-3104.



Tulsa Performance Scaling by Processor Count

Compares performance of Dual-Core Intel® Xeon® Processor MP 7140M ("Tulsa") with Intel E8501 chipset-based platform ("Truland") in one, two, and four processor configuration



Data Source: Intel internal measurement (July 2006). See backup for details.

INTEL Corporation

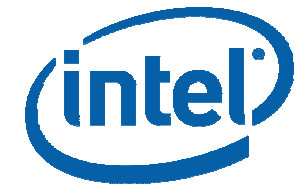
27 21 Aug 06 The Tulsa Processor
JDG

*Other names and brands may be claimed as property of others. System Configuration details and/or links to posted results in backup. Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit <http://www.intel.com/performance/resources/limits.htm> or call (U.S.) 1-800-628-8686 or 1-916-356-3104.



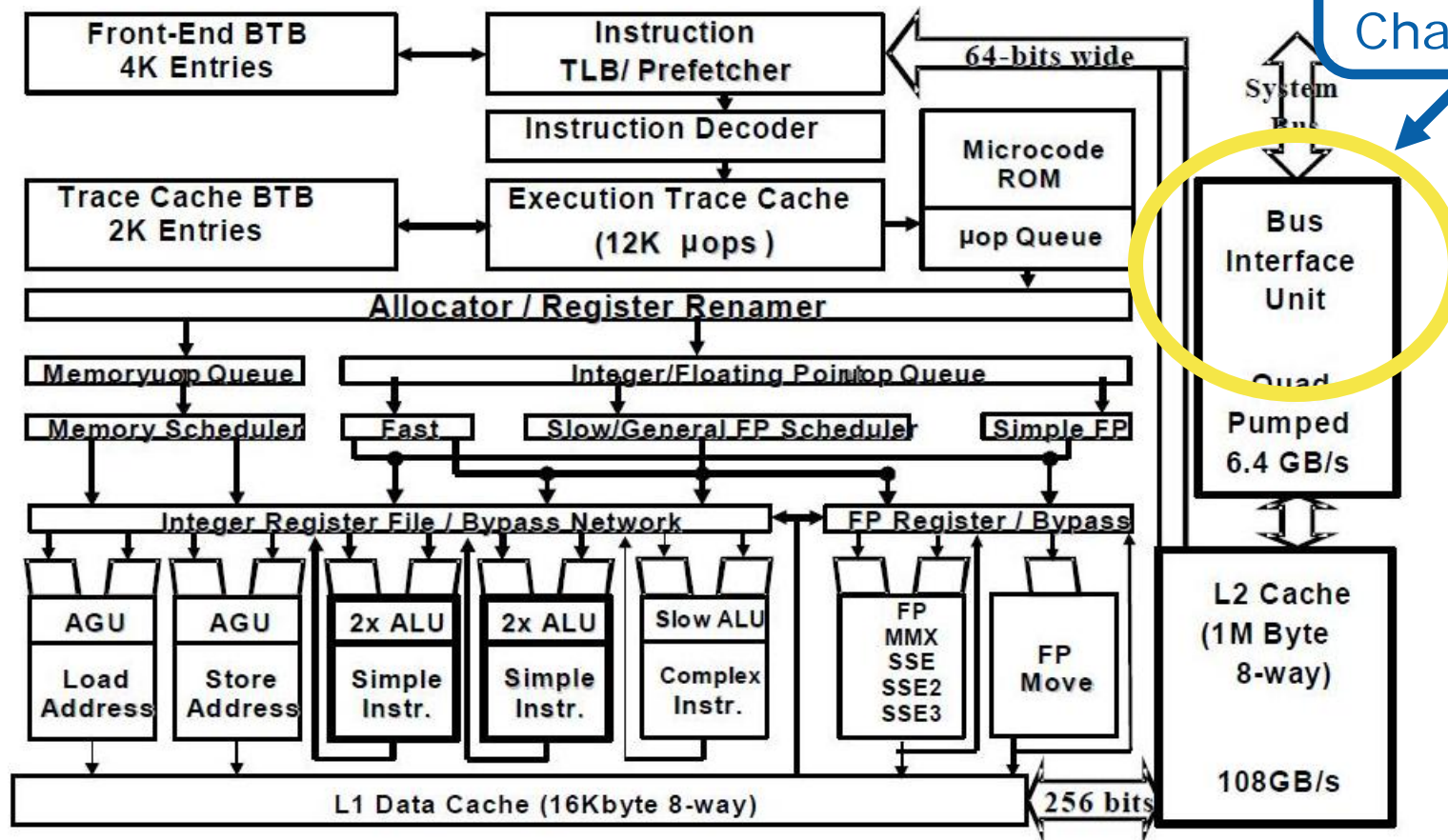
Conclusion

- An appropriate balance of cache and core resources can achieve platform performance levels far beyond a platform's original design targets
- A 2nd order cache benefit – queuing latency reduction afforded by fewer transactions – can be a significant contributor to overall memory latency reduction and thereby performance
- Morals of this team engineering story:
 - Robust and creative engineering can provide market-leading performance
 - As with some many things, the most difficult part of doing something is deciding – and committing – to do it
 - Compact teams can accomplish great things



Supplementary Slides

NetBurst Block Diagram



From <ftp://download.intel.com/technology/itj/2004/volume08issue01/vol8iss1.pdf>

Original Netburst Pipeline

Basic Pentium 4 Processor Misprediction Pipeline

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
TC Nxt IP		TC Fetch		Drive	Alloc		Rename	Que	Sch	Sch	Sch	Disp	Disp	RF	RF	Ex	Flgs	Br Ck	Drive

Note that 90 nm Netburst generation “extended the original Pentium 4 processor pipeline” to a “31-stage pipeline”. Intel has not made the specific pipeline changes public.

Bibliography

- ftp://download.intel.com/technology/itj/q12001/pdf/art_2.pdf
 - Original Intel Technology Journal Netburst microarchitecture paper
- <ftp://download.intel.com/technology/itj/2004/volume08issue01/vol8iss1.pdf>
 - Intel Technology Journal with updates to Netburst microarchitecture for the Prescott generation (the next generation, Cedar Mill, is quite similar)
- Tulsa at ISSCC 2006 (no on-line link yet)

Links to Posted Performance Results

These publicly pages present performance results for the Intel® Xeon® Processor MP 7140M (3.40 GHz, 800 MT/s FSB, 16 MB L3 cache) and describe the system configuration used to obtain the results.

- Compendium of published performance results
http://www.intel.com/performance/server/xeon_mp/index.htm
- SPECint*_rate_base_2000
<http://www.spec.org/cpu2000/results/res2006q3/cpu2000-20060807-06940.html>
- SAP* 2-tier SD SQL2005*
<http://www.sap.com/solutions/benchmark/index.epx>
- SPECfp*_rate_base_2000
<http://www.spec.org/osg/cpu2000/results/res2006q3/cpu2000-20060724-06782.html>
- SPECjbb* 2005
<http://www.spec.org/jbb2005/results/res2006q3/jbb2005-20060731-00160.html>



Performance System Configurations

Performance per Watt and Processor Count Performance scaling

SPECcpu2000 suite: Compute-intensive workload focusing on floating-point and integer speed and throughput. Performance estimates based on Intel internal measurement.

- Baseline Platform Configuration: Intel® SR4850HW4 Server System (Harwich with 800MT/s) using 4x Dual-Core Intel® Xeon® Processor MP 7041 (3.00 GHz, 800 MT/s FSB, 2x 2 MB L2 cache), HW/ADJSECT PREFETCH=ON, 8GB DDR2-400 (8x1GB PC2-3200R-333), Microsoft* Windows* Server 2003 Enterprise Edition SP1 32-bit, benchmark 1.3 using internally compiled Intel® C/C++ and Fortran Compiler version 9.1 for 32-bit.
- New Platform configuration: Intel® SR4850HW4 Server System (Harwich with 800MT/s) using 4x Dual-Core Intel® Xeon® Processor MP 7140M (3.40 GHz, 800 MT/s FSB, 16 MB L3 cache), HW/ADJSECT PREFETCH ON, 8GB DDR2-400 (8x1GB PC2-3200R-333), Microsoft* Windows* Server 2003 Enterprise Edition SP1 32-bit, benchmark 1.3 using internally compiled Intel® C/C++ and Fortran Compiler version 9.1 for 32-bit.



Performance System Configurations (cont'd)

Platform Performance, Performance per Watt, and Processor Count

Performance scaling

Database Performance: OLTP – On-Line Transaction Processing; represents the transaction throughput of a database server in a transaction processing client/server environment. The experiment measures the power and capacity of database software and server hardware using the transaction processing rate.

- Baseline Platform Configuration: Intel® SR6850HW4/M Server System (Harwich with 800MT/s) using 4x Dual-Core Intel® Xeon® Processor MP 7041 (3.00 GHz, 800 MT/s FSB, 2x 2 MB L2 cache), HW/ADJSECT PREFETCH=OFF, 64GB DDR2-400 (16x4GB PC2-3200R-333), Microsoft* Windows* Server 2003 Enterprise Edition SP1 x64.

Storage Configuration

- 854 15K RPM Seagate SCSI disks
 - 4 QLE2362 PCI-E QLogic Dual-port adapters
 - 1 QLA2342 PCI-X QLogic Dual-port adapters
- New Platform configuration: Intel® SR6850HW4/M Server System (Harwich with 800MT/s) using 4x Dual-Core Intel® Xeon® Processor MP 7140M (3.40 GHz, 800 MT/s FSB, 16 MB L3 cache), HW/ADJSECT PREFETCH=OFF, 64GB DDR2-400 (16x4GB PC2-3200R-333), Microsoft* Windows* Server 2003 Enterprise Edition SP1 x64.

Storage Configuration

- 994 15K RPM Seagate SCSI disks
- 3 QLA 2342 PCI-X QLogic Dual-port adapters
- 4 QLA 2362 PCI-E QLogic Dual-port adapters

Performance System Configurations (cont'd)

Processor Count Performance scaling

Enterprise Resource Planning on 2-tier: Workload emulates a Sales and Distribution application and helps ERP. Measured in number of concurrent users supported. Performance estimates based on Intel internal measurement.

- New Platform configuration: Intel® S3E3134 Server System (Harwich with 800MT/s) using 4x Dual-Core Intel® Xeon® Processor MP 7140M (3.40 GHz, 800 MT/s FSB, 16 MB L3 cache), HW/ADJSECT PREFETCH=OFF, 8GB DDR2-400 (8x1GB PC2-3200R), SuSE* LINUX* Enterprise 9 x86_64 SP2 2.6.5-191-smp, SAP* R/3 Enterprise ECC5.0 SR1 x86_64, Oracle9i* Enterprise Edition release 9.2.0.6.0 64-bit.



Performance System Configurations (cont'd)

Performance per Watt and Processor Count Performance scaling

SPECjbb*2005 v1.06: This workload evaluates the performance of Server-side Java Application. Measured in Operations Per Second. Performance estimates based on Intel internal measurement.

- Baseline Platform Configuration: Intel® SR4850HW4 Server System (Harwich with 800MT/s) using 4x Dual-Core Intel® Xeon® processor 7041 (3.00 GHz, 800 MT/s FSB, 2x 2 MB L2 cache), HW / ADJSECT PREFETCH=OFF, 16GB DDR2-400 (16x1GB PC2-3200R), Microsoft* Windows* Server 2003 Enterprise Edition x64 SP1, BEA* Internal JRockit* 5.0 64bit, large page enabled, 4 JVM instances.
- New Platform configuration: Intel® SR4850HW4 Server System (Harwich with 800MT/s) using 4x Dual-Core Intel® Xeon® processor 7140M (3.40 GHz, 800 MT/s FSB, 16 MB L3 cache), HW PREFETCH=OFF/ADJSECT PREFETCH=ON, 16GB DDR2-400 (16x1GB PC2-3200R), Microsoft* Windows* Server 2003 Enterprise Edition x64 SP1, BEA* Internal JRockit* 5.0 64bit, large page enabled, 4 JVM instances.



INTEL Corporation

38 21 Aug 06 The Tulsa Processor
JDG

